(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 February 2003 (13.02.2003)

PCT

(10) International Publication Number
WO 03/012713 A1

(51) International Patent Classification[7]: G06F 17/60

(21) International Application Number: PCT/NZ02/00138

(22) International Filing Date: 30 July 2002 (30.07.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

513265	30 July 2001 (30.07.2001)	NZ
516382	24 December 2001 (24.12.2001)	NZ
519832	27 June 2002 (27.06.2002)	NZ

(71) Applicant (*for all designated States except US*): COMPUDIGM INTERNATIONAL LIMITED [NZ/NZ]; Level 16, Compudigm House, 49 Boulcott Street, Wellington (NZ).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): **Cardno, Andrew John**, [NZ/NZ]; Level 16, Compudigm House, 49 Boulcott Street, Wellington (NZ). **Mulgan, Nicholas John**, [NZ/NZ]; Level 16, Compudigm House, 49 Boulcott Street, Wellington (NZ). **BEARD, Richard Thomas**, [NZ/NZ]; Level 16, Compudigm House, 49 Boulcott Street, Wellington (NZ).

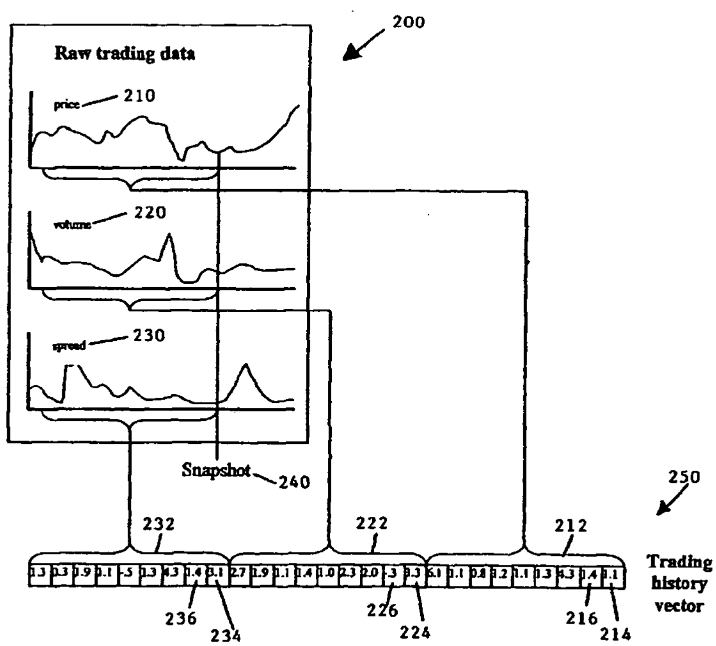
(74) Agents: **ADAMS, Matthew D** et al.; A J Park, 6th Floor Huddart Parker Building, Post Office Square, PO Box 949, 6015 Wellington (NZ).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: DATA VISUALISATION SYSTEM AND METHOD



(57) Abstract: The invention provides a data visualisation system comprising time-variant data maintained in computer memory; a time series analysis component configured to create one or more vectors from time-variant data; and a self-organising map component configured to generate and display a two-dimensional representation including one or more vector representations. The data visualisation system also includes a contour generator configured to generate and display one or more contour lines around each vector representation. The invention also provides a related method of data visualisation.

WO 03/012713 A1

ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

DATA VISUALISATION SYSTEM AND METHOD

FIELD OF INVENTION

The invention relates to a data visualisation system and method, particularly but not solely designed to identify patterns in time-variant data, for example stock exchange trading data. The invention is particularly suitable in identifying patterns in time-variant multivariate data.

BACKGROUND TO INVENTION

One of the biggest challenges in the financial markets is managing and interpreting the huge amounts of data and the limitation of existing tools to handle that data. Data associated with financial markets is usually time-variant. The price of stocks and shares is associated with a particular time interval, and significant price movements, both upward and downward can occur at different time intervals.

It is particularly important to identify patterns in price movements over time in order to predict short term future movements in the data. Such movements generally depend on several variables which include price, volume and spread and are particularly difficult to identify.

SUMMARY OF INVENTION

In broad terms in one form the invention comprises a data visualisation system comprising time-variant data maintained in computer memory; a time series analysis component configured to create one or more vectors from the time-variant data; a self-organising map component configured to generate and display a two-dimensional representation including one or more vector representations; and a contour generator configured to generate and display one or more contour lines around each vector representation.

In broad terms in another form the invention comprises a method of data visualisation comprising the steps of maintaining time-variant data in computer memory; creating one or more vectors from the time-variant data; generating and displaying a two-dimensional representation including one or more vector representations; and generating and displaying one or more contour lines around each vector representation.

BRIEF DESCRIPTION OF THE FIGURES

Preferred forms of the data visualisation system and method will now be described with reference to the accompanying figures in which:

Figure 1 shows a block form of a system in which one form of the invention may be implemented;

Figure 2 shows the operation of the time series analysis component of Figure 1;

Figure 3 shows a two-dimensional representation generated by the self-organising map component of Figure 1; and

Figures 4, 5 and 6 illustrate the application of the invention to different forms of time-variant data.

DETAILED DESCRIPTION OF PREFERRED FORMS

Figure 1 illustrates a block diagram of the preferred system 10 in which one form of the present invention may be implemented. The system could include one or more interactive client workstations 20 on which data can be displayed. The system 10 includes time-variant data 30 maintained in computer memory, for example historical market data representing stock exchange trading data. The data could also represent fixed interest/bonds, futures, options, foreign exchange or any other tradable instruments.

The time-variant data could be stored in a relational database or object-oriented database. Each stock trade is associated with a price value and a time value. The database 30 is configured to enable time-variant data to be retrieved from it. Examples of retrieved data could include a series of data values representing the price of a particular stock at different time intervals. The time-variant data could also include, for each trade, a trade price, a trade size, time of trade, a trade type (for example on market or over the counter), a current bid price and a current ask price.

The system 10 also includes a time series analysis component 40 configured to create one or more vectors from the time-variant data, as will be described below with reference to Figure 2. The time series analysis component is preferably implemented as a computer program installed and operating on a computing device. Alternatively, it is envisaged that the time series analysis component could be implemented in hardware form.

The system 10 further includes a self-organising map (SOM) component 50 which is configured to generate and display a two-dimensional representation which includes representations of one or more of the vectors obtained from the time series analysis component 40. The self-organising map component 50 is preferably a software program or hardware equivalent which implements one or more artificial neural networks to map one or more of the vectors obtained from the time series analysis 40 into two dimensions.

A self-organising map is a neural network or neural network combination in which the parameters that are trained are a set of cell vectors that exist in the same high dimensional space as the input data vectors. Each of these represents a transformation between one cell on the self-organising map and the high dimensional space. To train the self-organising map, each data vector is assigned to the cell whose vector is closest to it in the high dimensional space. The cell vectors are then iteratively moved around the high dimensional space so that each cell vector is moved towards the average of the vectors assigned to it and each cell vector is moved to a lesser extent towards the average of the vectors assigned to neighbouring or nearby cells. The effect of this is that the vectors

mirror the data and that neighbouring cells in the two-dimensional space are near each other in the higher dimensional space.

A contour generator or visualisation engine 60 is configured to generate and display one or more contour lines around each vector representation as will be described below with reference to Figure 3.

The system 10 optionally includes a cluster analysis component 70 to determine trends or properties in the output from the self-organising map component 50.

The output of this cluster analysis can be exported for further analysis if required, or the properties of a cluster can be used to define a leading key performance indicator or indicators 80.

Market data as a real time feed 90 is transmitted to a short term data store 100 and/or a dynamic template 110. Data is transmitted from the short term data store 100 and/or dynamic template 110 to a visualisation engine 120 and/or real time or near real time visualisation engine 130. The visualisation engine 120 is connected to a further interactive client 140 whereas the real time visualisation engine 130 could be configured to broadcast images to real time clients 150, for example 150A, 150B and 150C. Semi-static reference data, for example market structured data, could also provide input to the dynamic template 110A.

Figure 2 illustrates a sample vector created by the time series analysis component 40 of Figure 1. Raw trading data 200 is obtained from the historical market database 30 from Figure 1. The time-variant data could be multivariate data, having associated with it one or more variables. These variables could include, for example, the price of shares or stocks, the volume traded and the spread in prices. The time series analysis component obtains from the raw trading data values for the variables price 210, volume 220 and spread 230 at a particular time value or snapshot 240. Values of these variables at preceding time values are also obtained.

The resulting vector 250 represents a series of values for one or more variables, for example values 212 for price variable 210, values 222 for volume variable 220 and values 232 for spread variable 230.

Preferably, the number of values 212 equals the number of values 222 and 232. Furthermore, value 214, based on the value of price at a particular time value, is preferably sampled at the same time value for value 224 and value 234. Values 216, 226 and 236 are preferably all sampled at the same time value but at a preceding time value to the time value associated with values 214, 224 and 234.

Only three variables are shown in Figure 2 for simplicity, and it will be appreciated that the number of actual variables could be many times this number. Furthermore, only nine different time values are shown for simplicity and it will be appreciated that the number of time values in practice could be many times this number.

The resulting vector is a trading history vector in n-dimensional space where n can be in the hundreds or even thousands corresponding to the snapshot or current time value. By sampling thousands of such snapshots of different stocks at different times, the time series analysis component is able to create a large space of high-dimensional vectors. Within this space, there are likely to be significant structures, representing recurrent patterns in the trading data, but the vast dimensionality of the space makes it impossible for humans to discover these structures unaided.

Where necessary, the data in one or more of the resulting vectors is transformed, scaled and statistically normalised to ensure that comparisons between the often disparate variables are meaningful. Every dimension can be scaled to have a range from 0 to 1 – a continuous variable z has its values z_i scaled by $z_i \rightarrow (z_i - z_{min})/(z_{max} - z_{min})$ (where z_i are the values of a dimension z across the varying data points, indexed i; z_{min} (z_{max}) is the minimum (maximum) of the z_i).

A bistate or Boolean variable (male/female, true/false) is transformed to values 0,1 and a multi-state discrete variable with N+1 values is transformed to (1/N) {0,1,2,...,N}. More complicated order-preserving transformations, involving the expected or known variance, or higher order modes of the variable distribution can be used to normalise the variable to a uniform, normal or other standard distribution within the unit interval.

The vector or vectors resulting from the time series analysis component 40 is then passed to the self-organising map component 50. Figure 3 illustrates an example 2-dimensional representation generated by the self-organising map component. Each data point represents a representation of a vector created by the time series analysis component 40. The two-dimensional representation 300 for example includes data points 310, 320, 330 and 340. Preferably the mapping is performed in such a way that vectors which are similar in the high dimensional space remain close together in the 2-dimensional output. On this basis, vectors 310 and 320 are expected to be more similar in high dimensional space due to their close proximity in the representation 30, whereas vectors 310 and 340 are expected to be more dissimilar due to the spacing between these vectors in the representation 300.

The visualisation engine 60 and/or the self-organising map component 50 is configured to calculate, for each vector, the price movement occurring in a given time interval after the snapshot or current time associated with each vector. The absolute value of price movement is a numerical value which is associated with the individual data points representing individual vectors. Vectors 310 and 320 have no associated price movement and are shown as such in the representation. Vector 330 has some price movement and vector 340 has a larger price movement.

The visualisation engine/contour generator treats the price movements associated with each data point as a maximum value and generates and displays one or more contour lines around each data point representing a vector. Each contour line represents data values which are less than the price movement associated with the data point around which the contour line is displayed. The preferred form technique for contouring is described in

patent specification WO 00/77682 to Compudigm International Limited entitled "Data Visualisation System and Method".

The use of contouring in this manner enables a viewer to rapidly determine patterns which precede consistent price movements. If certain clusters show strong subsequent price movements for all the trading history vector within, then that cluster represents a trading pattern that is a powerful predictive tool.

The viewer may then use a cluster analysis tool 70 to determine the properties of a particular cluster under consideration, resulting in a formula that can produce, for any trading history vector, a value that represents how strongly that vector would belong to the selective cluster.

For each cluster that shows strong subsequent price movements, a leading indicator is defined. For example, the Euclidean distance between a new data point and the centre of the cluster in the original high dimensional space produces a value that can be plotted. Alternatives include setting thresholds on each dimension or more sophisticated techniques such as analysis of variance (ANOVA).

Referring to Figure 1, as live trading data 90 arrives from a data feed, the recent trading history of each stock can be turned into history vectors. By applying the leading indicator definition accordingly, the system can assign each stock a number indicating the tendency of that stock to move up or down in price in the near future. By plotting this leading indicator in real time, either on traditional charts or on a contoured visualisation of the entire market, traders will gain a valuable tool to assist their trading decisions.

The invention has several advantages. It does not attempt to replace human intuition, like other artificial neural network-based trading tools, but instead amplifies intuition with visualisations. The contours show patterns more strongly than other visualisation techniques. The invention combines multiple variables into a single pattern represented

as the history vector. The invention can also be adapted to work on many different sorts of time-variant data.

The invention provides an advance upon existing technology as it is a unique hybrid of time series transformation, self-organising maps, statistical clustering and contouring techniques. The invention is also able to transform time-variant multivariate data into history vectors suitable for input to the self-organising map. The invention is also interactive, enabling analysts to experiment with variable selection and weightings.

It is envisaged that the invention could be applied to any time-variant data in addition to financial market data. In one form the time-variant data could include revenues from gaming machines in a casino context. Referring to Figure 4, machines are grouped by similarity into clusters of a particular nature, in this case denomination, and sub-clusters for example spend patterns. Onto these groupings, a contoured representation identifying a key performance indicator, for example turnover, is overlaid to show performance of groupings based on the customer base of the organisation or a segment of interest within it.

Figure 5 illustrates the organisation of time-variant data into a series of layers including mesh block cluster boundary layers or mesh block data layers. This display could also include a contour layer of market opportunity or any key performance indicator selected for each mesh block, and a thematic layer showing particular mesh blocks of interest, for example high value, low value, people that buy a particular product, or any useful profile per mesh block. The projection could also include a selection layer which indicates the newly selected mesh blocks.

Figure 6 illustrates the invention applied to data in order to provide knowledge management. The knowledge management application relies on starting with a corpus (group) of documents. The documents are put through a lexical analyser system which finds a lexicon comprising all the non-trivial words (or stems or lexical roots of words) in the corpus and assigns a value to each {document, word/stem} pair. The value of each

pair is the importance of that word in that document, e.g., 0 if the word does not appear, Maximum if it is in the title and occurs in every paragraph. Maximum if the word is the distinguishing (from the remainder of the corpus) topic of the document.

Typically words that have a high rating in nearly all documents (company name) are ignored as their discriminating power is limited. The set of values for one document – one for each word in the lexicon - constitutes the vector for that document that is fed to the self-organising map. The full input is one vector for each document. Note that each dimension is of the same type as they all measure the same type of relationship.

The foregoing describes the invention including preferred forms thereof. Alterations and modifications as will be obvious to those skilled in the art are intended to be incorporated within the scope hereof, as defined by the accompanying claims.

CLAIMS:

1. A data visualisation system comprising:

time-variant data maintained in computer memory;

a time series analysis component configured to create one or more vectors from the time-variant data;

a self-organising map component configured to generate and display a two-dimensional representation including one or more vector representations; and

a contour generator configured to generate and display one or more contour lines around each vector representation.

2. A data visualisation system as claimed in claim 1 wherein the time series analysis component is configured to create at least one vector based on the value of a variable at a time value, and a series of values of the variable at preceding time values.

3. A data visualisation system as claimed in claim 2 wherein the time series analysis component is configured to create at least one vector based on the values of a plurality of variables at a time value, and a series of values of the variable at preceding time values.

4. A method of data visualisation comprising the steps of:

maintaining time-variant data in computer memory;

creating one or more vectors from the time-variant data;

generating and displaying a two-dimensional representation including one or more vector representations; and

generating and displaying one or more contour lines around each vector representation.

5. A method of data visualisation as claimed in claim 4 wherein the step of creating one or more vectors is based on the value of a variable at a time value, and a series of values of the variable at preceding time values.

6. A method of data visualisation as claimed in claim 5 further comprising the step of creating at least one vector based on the values of a plurality of variables at a time value, and a series of values of the variable at preceding time values.

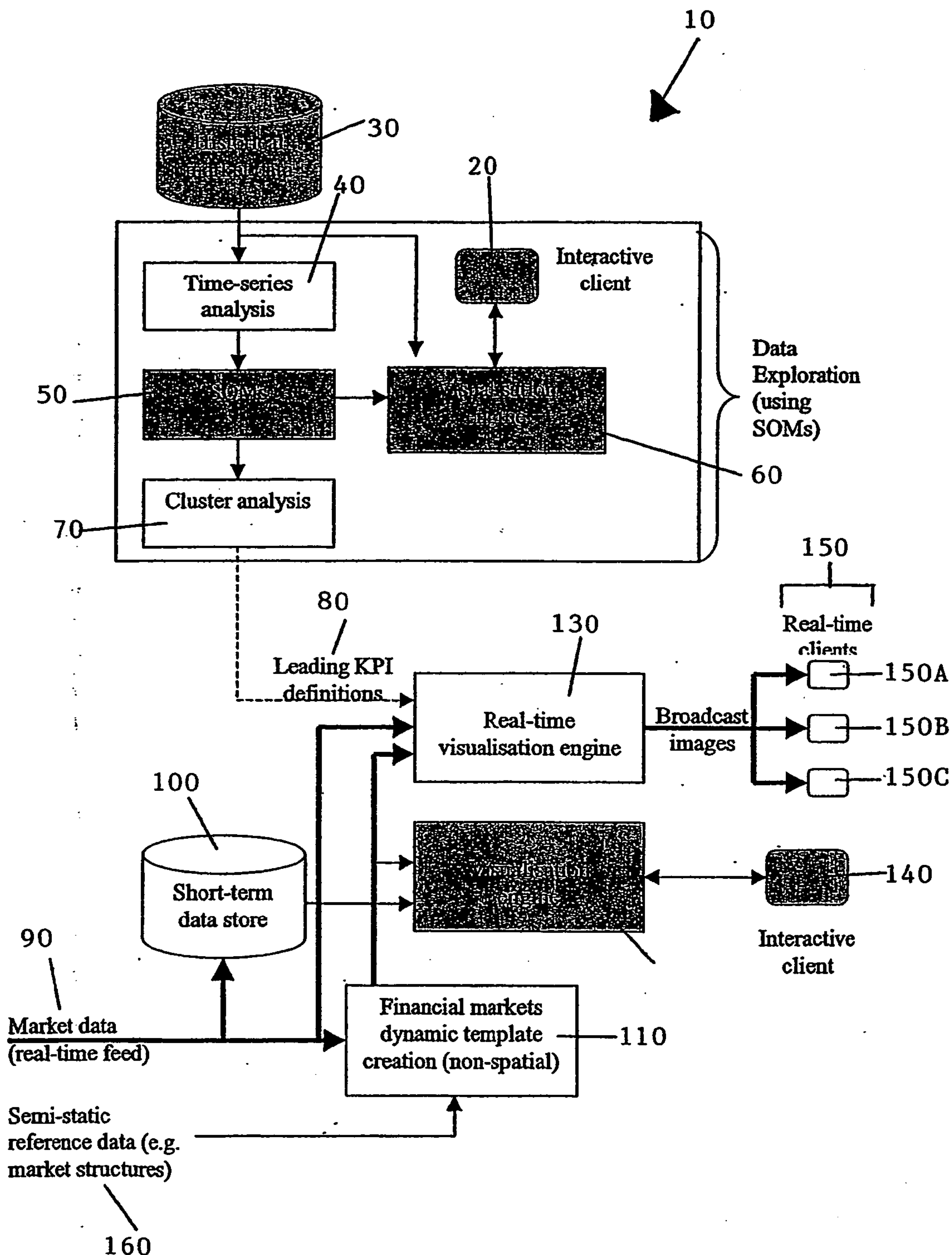

10
30
20
40
Interactive client
Time-series analysis
50
Data Exploration (using SOMs)
60
Cluster analysis
70
150
80
130
Real-time clients
Leading KPI definitions
150A
Real-time visualisation engine
Broadcast images
150B
150C
100
140
Short-term data store
Interactive client
90
Market data (real-time feed)
Financial markets dynamic template creation (non-spatial)
110
Semi-static reference data (e.g. market structures)
160

FIGURE 1

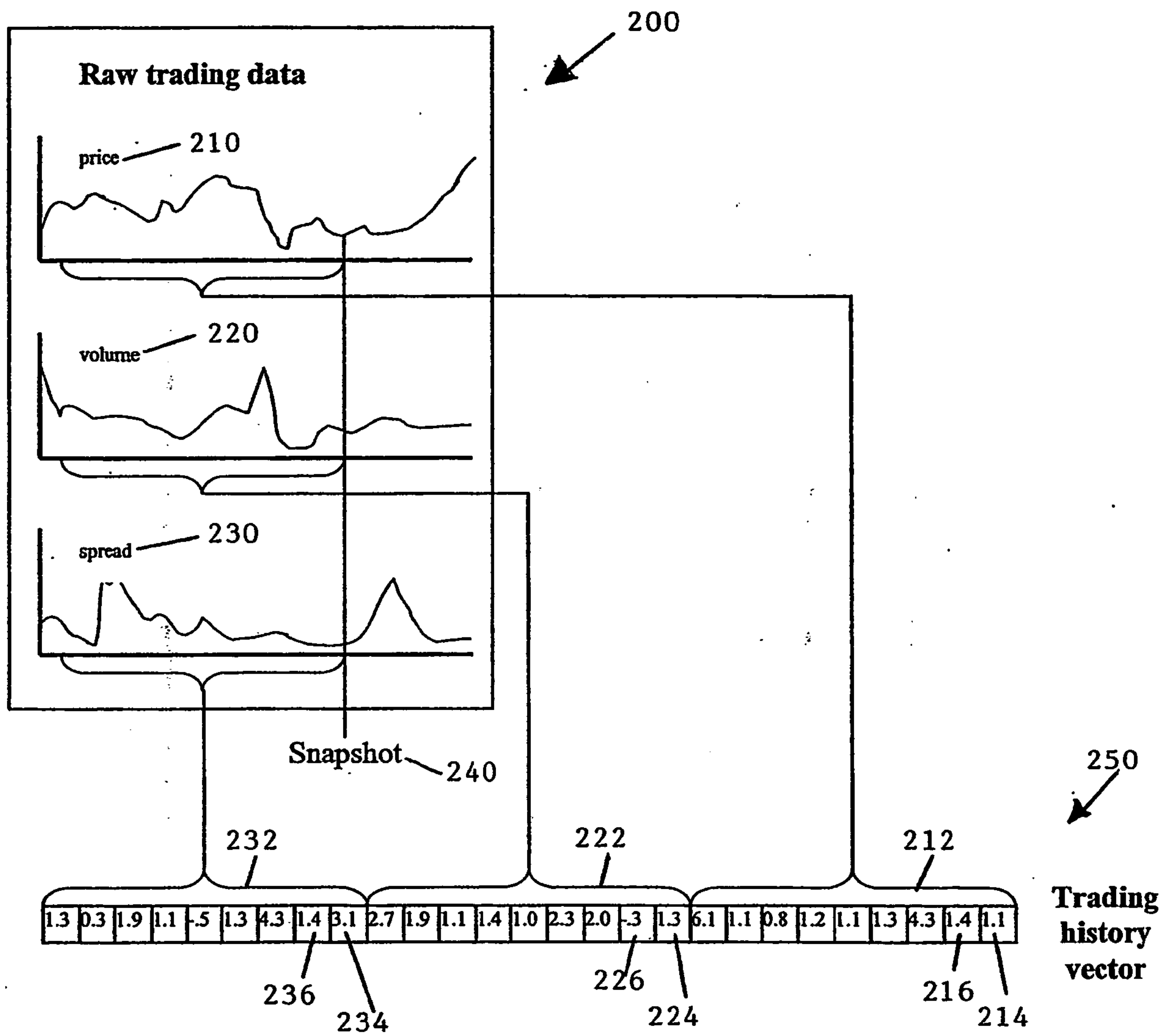

200
Raw trading data
price 210
volume 220
spread 230
Snapshot 240
250
232
222
212
1.3 0.3 1.9 1.1 -.5 1.3 4.3 1.4 3.1 2.7 1.9 1.1 1.4 1.0 2.3 2.0 -.3 1.3 6.1 1.1 0.8 1.2 1.1 1.3 4.3 1.4 1.1
Trading history vector
236
234
226
224
216
214

FIGURE 2

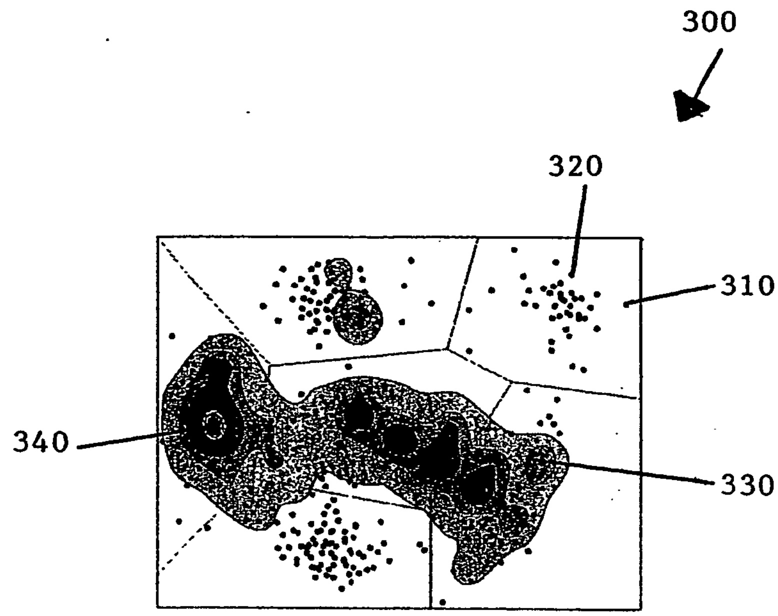


FIGURE 3

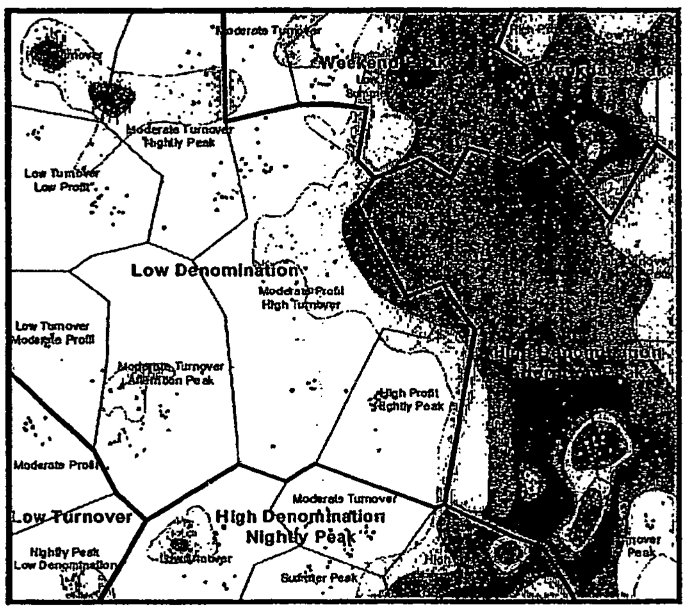

Low Denomination
Moderate Turnover
Nightly Peak
Low Turnover
Low Profit
Moderate Profit
High Turnover
Low Turnover
Moderate Profit
Moderate Turnover
Afternoon Peak
High Profit
Nightly Peak
Moderate Profit
Low Turnover
Nightly Peak
Low Denomination
High Denomination
Nightly Peak
Moderate Turnover
Summer Peak
Weekend Peak

FIGURE 4

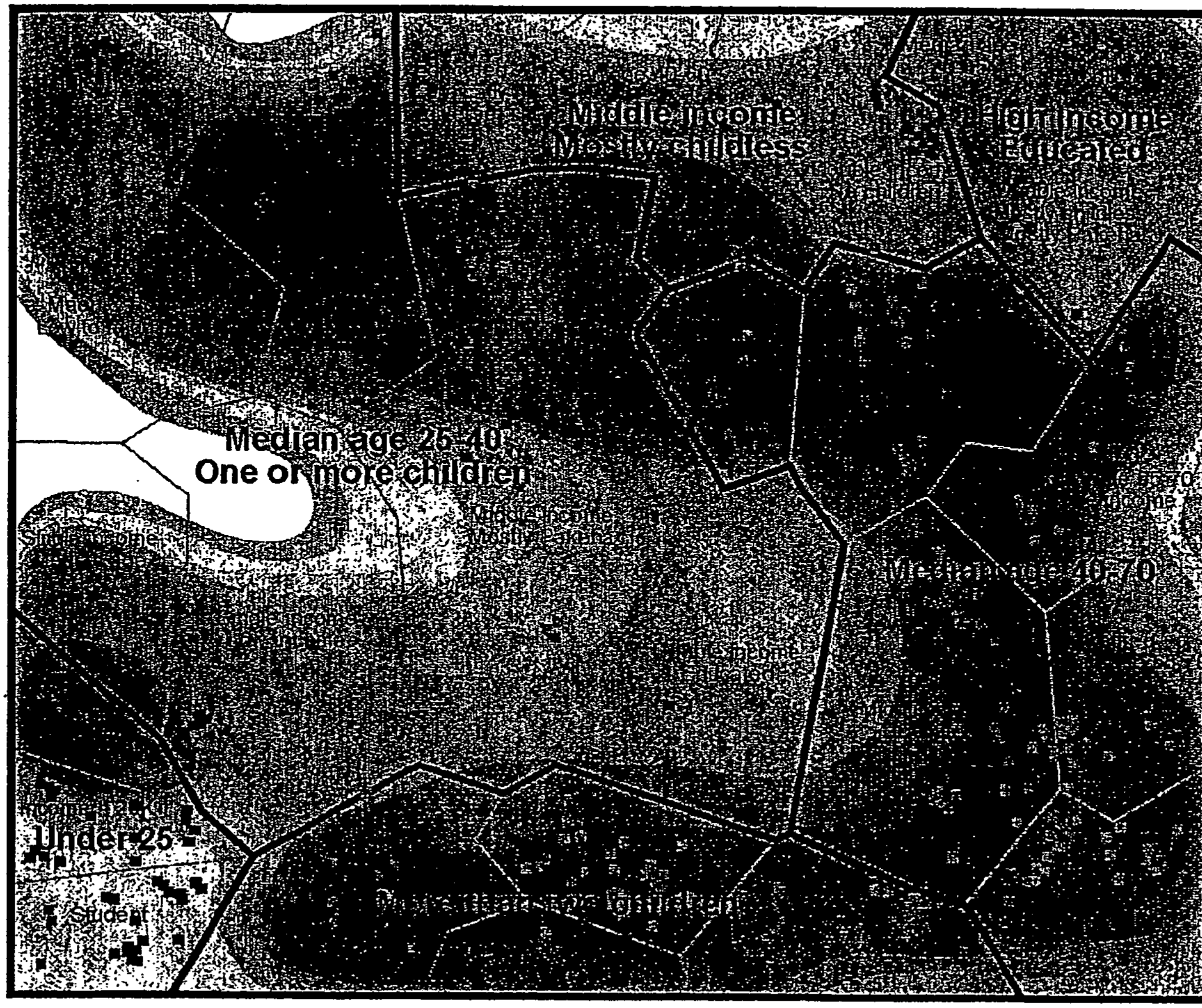

Middle income
Mostly childless
High income
Educated
Median age 25-40
One or more children
Median age 40-70
Under 25
Student

FIGURE 5

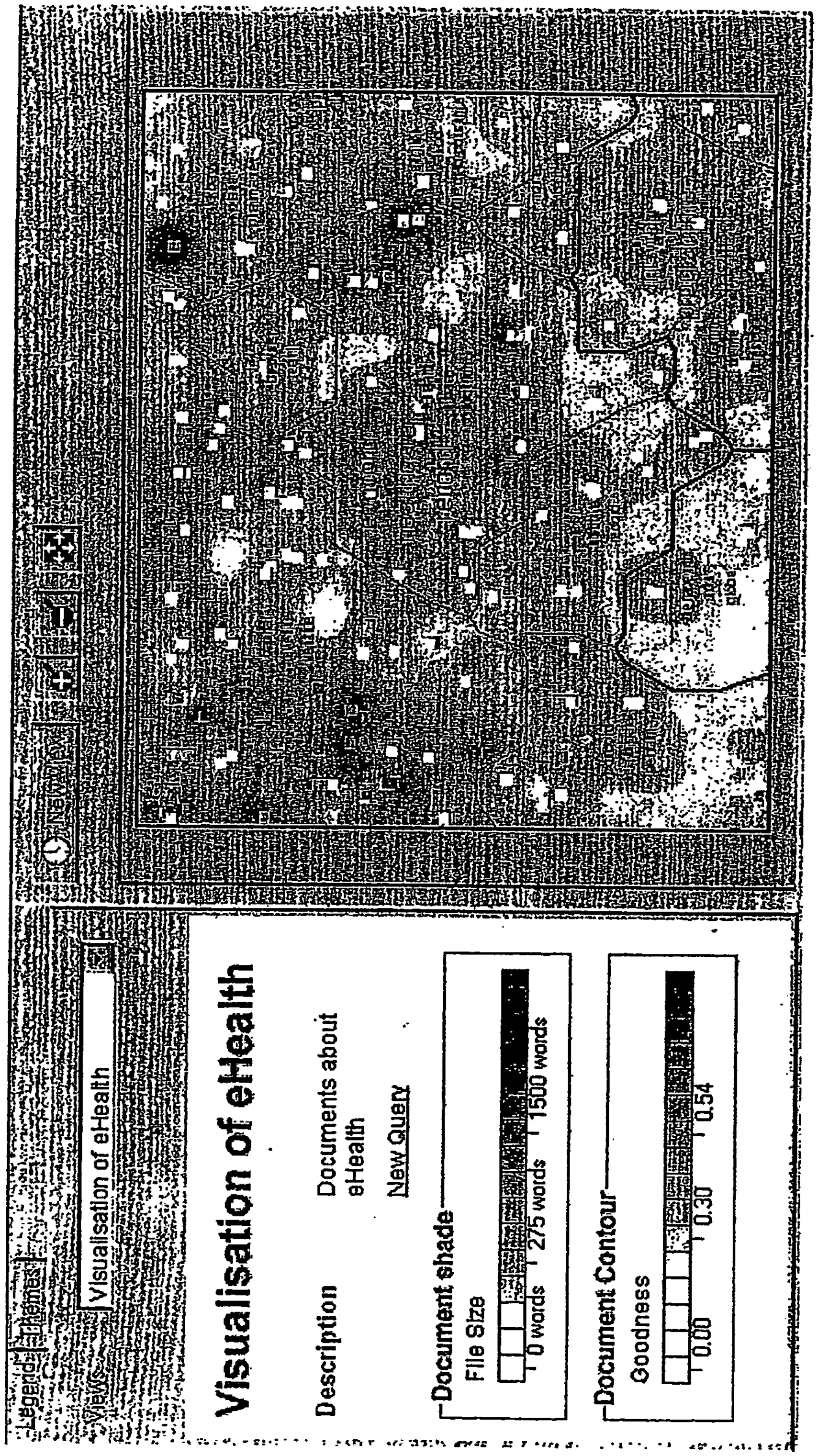

Legend
Themes
Views
Visualisation of eHealth
Visualisation of eHealth
Description
Documents about eHealth
New Query
Document shade
File Size
0 words
275 words
1500 words
Document Contour
Goodness
0.00
0.30
0.54

FIGURE 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/NZ02/00138

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. [7]: G06F 17/60

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WPAT, USPTO Web Patent Database, Esp@cenet, "visual, neural, self organising map, vector, contour etc."

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6035057 A (HOFFMAN) 7 March 2000 Entire document.	1-6
Y	WO 01/04808 A (COMPUDIGM INTERNATIONAL LIMITED) 18 January 2001 Entire document	1-6
A	US 6301579 B1 (BECKER) 9 October 2001	1-6
A	US 6307500 B1 (CORNMAN et al.) 23 October 2001	1-6

☐ Further documents are listed in the continuation of Box C ☒ See patent family annex

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search 8 October 2002	Date of mailing of the international search report 10 OCT 2002
Name and mailing address of the ISA/AU AUSTRALIAN PATENT OFFICE PO BOX 200, WODEN ACT 2606, AUSTRALIA E-mail address: pct@ipaustralia.gov.au Facsimile No. (02) 6285 3929	Authorized officer **P. THONG** Telephone No : (02) 6283 2128

This Annex lists the known "A" publication level patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

Patent Document Cited in Search Report		Patent Family Member			
US	6035057	CA	2199588	US	6278799
WO	200104808	AU	200063247	EP	1212716
US	6301579	NONE			
US	6307500	NONE			

END OF ANNEX